Psychological Methods
2016, Vol. 21, No. 1, 47–60


1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000036

# Planning Sample Sizes When Effect Sizes Are Uncertain: The Power-Calibrated Effect Size Approach

Blakeley B. McShane and Ulf Böckenholt
Northwestern University

Statistical power and thus the sample size required to achieve some desired level of power depend on the size of the effect of interest. However, effect sizes are seldom known exactly in psychological research. Instead, researchers often possess an estimate of an effect size as well as a measure of its uncertainty (e.g., a standard error or confidence interval). Previous proposals for planning sample sizes either ignore this uncertainty thereby resulting in sample sizes that are too small and thus power that is lower than the desired level or overstate the impact of this uncertainty thereby resulting in sample sizes that are too large and thus power that is higher than the desired level. We propose a power-calibrated effect size (PCES) approach to sample size planning that accounts for the uncertainty associated with an effect size estimate in a properly calibrated manner: sample sizes determined on the basis of the PCES are neither too small nor too large and thus provide the desired level of power. We derive the PCES for comparisons of independent and dependent means, comparisons of independent and dependent proportions, and tests of correlation coefficients. We also provide a tutorial on setting sample sizes for a replication study using data from prior studies and discuss an easy-to-use website and code that implement our PCES approach to sample size planning.

*Keywords:* power, sample size, effect size, statistical significance

*Supplemental materials:* http://dx.doi.org/10.1037/met0000036.supp



Recent difficulties in replicating prior results (Brodeur, Le, Sangnier, & Zylberberg, 2012; Francis, 2013; Ioannidis, 2005; Yong, 2012) have led to an increased interest in replication, study-planning, and research practices (Asendorpf et al., 2013; Brandt et al., 2014; Pashler & Wagenmakers, 2012). A particular area of focus has been on sample size considerations. Textbooks recommend setting sample sizes to achieve some prespecified level of power (typically 80%; Cohen, 1992). However, this recommendation can be difficult to implement in practice because "we never know true power . . . because we do not know the true effect size" (Cumming, 2014). This is true even for replication studies because prior data can at best provide an estimate of the effect size as well as an estimate of its error.

How should researchers attempting replication determine sample sizes in this setting, namely when they possess an estimate of the effect size as well as a measure of its uncertainty (e.g., a standard error or confidence interval)? We propose a power-calibrated effect size (PCES) approach to sample size planning that accounts for the uncertainty associated with estimates of effect sizes and yields sample sizes that reflect this uncertainty while providing the desired level of power. The PCES approach is based on the notion of expected power according to which power, when averaged over all possible effect sizes, is set to the desired level (e.g., 80%).

The PCES approach to sample size planning is attractive compared to alternative approaches because it accounts for uncertainty in a principled manner, that is, it is properly calibrated so that sample sizes determined using the PCES approach provide the desired level of power on average. Consequently, researchers who use the PCES approach are less likely to waste resources by setting sample sizes either too small or too large and thus having lower or higher than the desired level of power respectively. Because the PCES approach is properly calibrated, sample sizes based on it can be substantially smaller than those derived from alternative approaches that attempt to account for uncertainty.

Another principal benefit of the PCES approach is that it is easy to implement. The only modification to current practice required is that researchers must first compute the PCES; they may then continue using whatever techniques for sample size planning they currently use (e.g., textbook sample size formulae or statistical software). To facilitate this, we have developed an easy-to-use website that implements the PCES approach; the principal R (R Core Team, 2012) code underlying the website as well as R code to replicate analyses conducted in this paper are available both in the online supplementary materials and at the website and can be used without recourse to it. Because the PCES approach requires only a minor modification of current practice, it can be readily adopted by researchers interested in conducting studies that are adequately powered even when an effect size cannot be specified with a high level of accuracy.

In the remainder of this paper, we discuss the notion of expected power and derive the PCES. We show that the PCES properly

This article was published Online First December 14, 2015.

Blakeley B. McShane and Ulf Böckenholt, Marketing Department, Kellogg School of Management, Northwestern University.

Correspondence concerning this article should be addressed to Blakeley B. McShane, Marketing Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208. E-mail: b-mcshane@kellogg.northwestern.edu

Table 1
*Textbook Sample Size Formulae for Common Tests*

| Test | General | One-sided ($\alpha = 0.05$, $\beta = 0.20$) | Two-sided ($\alpha = 0.05$, $\beta = 0.20$) |
|---|---|---|---|
| Comparison of two independent means | $\frac{2\sigma^2(z_{1-\alpha} - z_\beta)^2}{\Delta^2}$ | $\frac{12.37\sigma^2}{\Delta^2}$ | $\frac{15.70\sigma^2}{\Delta^2}$ |
| Comparison of two dependent means | $\frac{\sigma_D^2(z_{1-\alpha} - z_\beta)^2}{\Delta^2}$ | $\frac{6.18\sigma_D^2}{\Delta^2}$ | $\frac{7.85\sigma_D^2}{\Delta^2}$ |
| Comparison of two independent proportions | $\frac{2\bar{p}(1-\bar{p})(z_{1-\alpha} - z_\beta)^2}{\Delta^2}$ | $\frac{12.37\bar{p}(1-\bar{p})}{\Delta^2} \leq \frac{3.09}{\Delta^2}$ | $\frac{15.70\bar{p}(1-\bar{p})}{\Delta^2} \leq \frac{3.92}{\Delta^2}$ |
| Comparison of two dependent proportions | $\frac{(z_{1-\alpha} - z_\beta)^2(p_{01} + p_{10})}{(p_{01} - p_{10})^2}$ | $\frac{6.18(p_{01} + p_{10})}{(p_{01} - p_{10})^2}$ | $\frac{7.85(p_{01} + p_{10})}{(p_{01} - p_{10})^2}$ |
| | $= \frac{(z_{1-\alpha} - z_\beta)^2}{4\left(\ddot{p} - \frac{1}{2}\right)^2(p_{01} + p_{10})}$ | $= \frac{1.55}{\left(\ddot{p} - \frac{1}{2}\right)^2(p_{01} + p_{10})}$ | $= \frac{1.96}{\left(\ddot{p} - \frac{1}{2}\right)^2(p_{01} + p_{10})}$ |
| Test of a correlation coefficient | $\frac{(z_{1-\alpha} - z_\beta)^2}{Z_\rho^2} + 3$ | $\frac{6.18}{Z_\rho^2} + 3$ | $\frac{7.85}{Z_\rho^2} + 3$ |

*Note.* The formulae give the sample size per condition for between-subjects designs and the total sample size for within-subjects designs and correlations. The null hypothesis is that of no difference or no correlation respectively. For two-sided tests, use $z_{1-\alpha/2}$ in place of $z_{1-\alpha}$ where $z_\gamma$ is the $100\gamma$ percentile of the standard normal distribution. In the table, $\Delta$ is the researcher's point estimate of the difference in means or proportions between the conditions, $\bar{p}$ is the researcher's point estimate of the average proportion between the conditions (i.e., $\bar{p} = (p_1 + p_2)/2$where $p_i$ is the researcher's point estimate of the proportion in condition $i$), $\ddot{p}$ is the researcher's point estimate of the change proportion (i.e., $\ddot{p} = p_{10}/(p_{01} + p_{10})$ where $p_{ij}$ is the researcher's point estimate of the proportion who selected $i$ then $j$), $Z_\rho$ is the Fisher $Z$-transformation of the researcher's point estimate of the correlation coefficient (i.e., $Z_\rho = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) = \text{arctanh}(\rho)$), $\sigma^2$ is the variance of the individual-level observations, and $\sigma_D^2$ is the variance of the individual-level differences. Our formulae assume without loss of generality that $\Delta$ and $\left(\ddot{p} - \frac{1}{2}\right)$ are positive; if $Z_\rho$ is negative use $|Z_\rho|$ in place of it. The formulae appear in Hays (1963) and similar textbooks.

calibrates power to the desired level while alternative approaches yield power that is lower or higher than the desired level. We then provide a tutorial on setting sample sizes for a replication study using data from prior studies. Next, we discuss our website and code, show how to use the website to reproduce the replication study results, and provide some additional examples. Finally, we conclude with recommendations.

## Expected Power and the PCES

In this section, we show how to account for the uncertainty associated with estimates of effect sizes. Previous proposals in the literature either ignore this uncertainty or overstate its impact. Our approach calibrates the effect size in a manner that reflects this uncertainty and provides the desired level of power on average.

Consider a researcher who is interested in some parameter $\theta$. In particular, suppose the researcher wants to test the null hypothesis $H_0 : \theta \leq 0$ versus the alternative hypothesis $H_1 : \theta > 0$ (or more generally $H_0 : \theta \in \Theta_0$ versus the alternative hypothesis $H_1 : \theta \in \Theta_1 = \Theta_0^c$ for $\theta \in \Theta = \Theta_0 \cup \Theta_1$). The researcher selects a hypothesis test, which is typically specified in terms of (a) a test statistic $T(X_1, \ldots, X_n) = T(\boldsymbol{X})$ that is some function of the $n$ sample data points and (b) a critical (or rejection) region R. The critical region R is typically selected so that, if the null hypothesis $H_0$ is true, the probability the test statistic $T(\boldsymbol{X})$ lies in the critical region R is no more than $\alpha$, the size of the test (i.e., the maximum probability of a Type I error or the minimum significance level). The null hypothesis $H_0$ is rejected when the test statistic $T(\boldsymbol{X})$ lies in the critical region R. For example, a researcher may be interested in testing whether $\mu$, the mean of $n$ independent and identically distributed normal random variables with known variance $\sigma^2$, is less than or equal to zero versus greater than zero (i.e., $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$) at $\alpha = 0.05$. In this case, the researcher might specify $T(\boldsymbol{X}) = \bar{X}/(\sigma^2/\sqrt{n})$ (i.e., the sample mean of the data divided by its standard error) as the test statistic and $T(\boldsymbol{X}) \geq 1.64$ as the critical region; if the null hypothesis is true, there is at most a 0.05 probability that $T(\boldsymbol{X}) \geq 1.64$.

Power is the probability that the random variable $T(\boldsymbol{X})$ lies in the critical region when the null hypothesis $H_0$ is false; thus, power is the probability of correctly rejecting the null hypothesis. The sampling distribution of $T(\boldsymbol{X})$ and thus power typically depend on both $\theta$ and $n$ while the critical region depends on $\alpha$ and can depend on $n$. If we write the probability density function of the sampling distribution of $T(\boldsymbol{X})$ as $f_T(\mathbf{x}|\theta, n)$ and the critical region as $R_{\alpha,n}$, then power $P(\theta, n, \alpha)$ is defined as

$$P(\theta, n, \alpha) = \mathbb{P}(T(\boldsymbol{X}) \in R_{\alpha,n}) = \int_{\mathbf{x}:T(\mathbf{x})\in R_{\alpha,n}} f_T(\mathbf{x}\,|\,\theta, n)d\mathbf{x} \quad (1)$$

where the integral is taken over the set of points $\mathbf{x}$ in the sample space of $\boldsymbol{X}$ such that the test statistic $T(\boldsymbol{X})$ lies in the critical region $R_{\alpha,n}$.

Best practices dictate setting the sample size $n$ to achieve adequate power where 80% is typically deemed adequate (Cohen, 1992). Thus, conditional on the parameter value $\theta$ and the size of the test $\alpha$ (typically $\alpha = 0.05$), researchers use Equation 1 to find the smallest $n$ that yields the desired level of power $1 - \beta$ (typically $\beta = 0.20$ so that power is 80%). In practice, this is typically accomplished either through textbook sample size formulae (e.g., those in Table 1) or statistical software (e.g., G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), R, SAS, SPSS, or Stata):

researchers plug the parameter value $\theta$, the size of the test $\alpha$, and the desired level of power into textbook formulae or software and obtain the sample size that yields the desired level of power.

This approach suffers from a serious limitation: it requires researchers to know $\theta$, which they of course do not (otherwise they would not need to collect data in the first place). However, in replication studies, researchers can use data from prior studies to form beliefs about (or estimate) $\theta$. These beliefs can often, at least loosely, be characterized as a probability density function $\pi(\theta|D)$ where conditioning on $D$ makes explicit the role of data $D$ from prior studies. Using $\pi(\theta|D)$, researchers then typically select a plausible value $\tilde{\theta}$ for $\theta$ and use this in conjunction with textbook formulae or software (i.e., as discussed in the prior paragraph).

A popular approach, which we term the point estimate approach, is to select as $\tilde{\theta}$ the mean, $\mathbb{E}[\theta] = \int_\Theta \theta\pi(\theta|D)\, d\theta$. While this seems like a reasonable value, in practice it typically results in power that is below the desired level on average. To understand this, consider the left panel of Figure 1, which illustrates the well-known fact that the secant line of a concave function lies below the graph of the function; since the secant line is a weighted average of the concave function $w\varphi(x_1) + (1 - w)\varphi(x_2)$ and the graph is the concave function of the weighted average $\varphi(wx_1 + (1 - w)x_2)$, we have $w\varphi(x_1) + (1 - w)\varphi(x_2) \leq \varphi(wx_1 + (1 - w)x_2)$. Jensen's inequality (Jensen, 1906) generalizes this fact to an arbitrary weighted average of an arbitrary number of points: if $X$ is a random variable and $\varphi$ is a concave function, then $\mathbb{E}[\varphi(X)] \leq \varphi(\mathbb{E}[X])$. Power $P(\theta, n, \alpha)$ is concave at least locally at sufficiently high values (i.e., those that are desirable in practice) because it asymptotes to one as a function of the sample size. Consequently, Jensen's inequality implies $\mathbb{E}[P(\theta, n, \alpha)] \leq P(\mathbb{E}[\theta], n, \alpha)$ as illustrated in the right panel of Figure 1. In sum, even if $\pi(\theta|D)$, the researcher's beliefs about $\theta$ based on prior data, is on average centered around the true value, the point estimate approach results in sample sizes that are too small and thus power that is below the desired level on average; this effect can be substantial in practice.

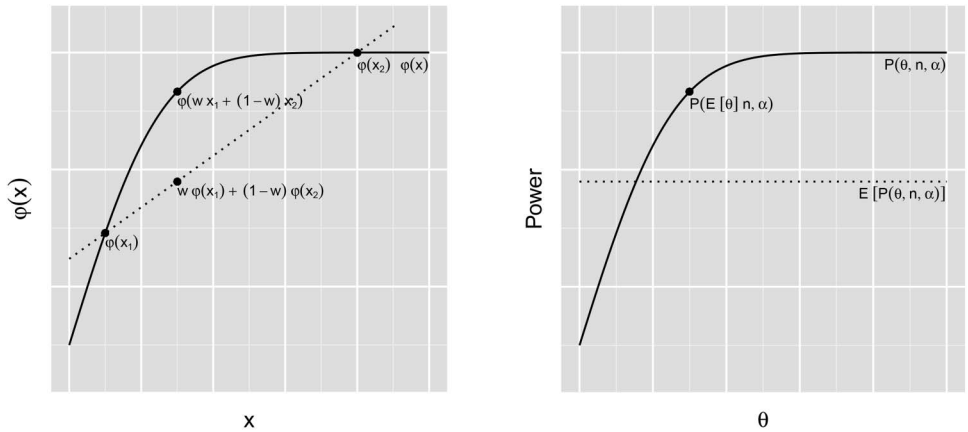
The shortcoming of the point estimate approach derives from the fact that it ignores the uncertainty about $\theta$ implied by $\pi(\theta|D)$: it treats $\theta$ as if it were known to be its mean. The safeguard power approach (Perugini, Gallucci, & Costantini, 2014) was developed to overcome this shortcoming. It recognizes the uncertainty implied by $\pi(\theta|D)$ and tries to account for it by choosing $\tilde{\theta} = Q(p)$ where Q is the quantile function corresponding to $\pi$ and $p$ is relatively small. Perugini et al. (2014) recommend $p = 0.20$ (i.e., the twentieth percentile), so that 80% of the time $\tilde{\theta}$ will be larger than $\theta$. While the safeguard power approach is an improvement over the point estimate approach in that it recognizes the uncertainty about $\theta$ implied by $\pi(\theta|D)$, it suffers from one major shortcoming: it does not properly calibrate power. In fact, the safeguard power approach is typically quite pessimistic: $\tilde{\theta} = Q(0.20)$ is too small resulting in sample sizes that are too large and power that is above, often substantially, the desired level (see Figure 1 of Perugini et al., 2014).

The limitation of each of these approaches is not in picking a value $\tilde{\theta}$ for $\theta$ *per se* but rather in picking that value indiscriminately (i.e., always picking the mean in the case of the point estimate approach and always picking the twentieth percentile in the case of the safeguard power approach). Given that $\pi(\theta|D)$ implies that some values of $\theta$ are relatively more likely while other values are relatively less likely, power should be calculated (and thus sample sizes set) in a manner that respects this variation in likelihood. This is known as the expected power approach to power calculation (Gillett, 1994; O'Hagan, Stevens, & Campbell, 2005; Spiegelhalter, Abrams, & Myles, 2004) and expected power $EP(\pi, n, \alpha)$ is defined as

$$EP(\pi, n, \alpha) = \int_\Theta P(\theta, n, \alpha)\pi(\theta|D)d\theta. \quad (2)$$

Thus, expected power is a weighted average of power $P(\theta, n, \alpha)$ where the weights correspond to the researcher's beliefs $\pi(\theta|D)$.

When $\theta$ is known, $\pi(\theta|D)$ is a probability mass function that puts probability one on the known value of $\theta$ and probability zero




(a) Concave Function and Secant Line (b) Power and Expected Power

*Figure 1.* Jensen's inequality and power. The left panel illustrates that the secant line of a concave function lies below the graph of the function, $w\varphi(x_1) + (1 - w)\varphi(x_2) \leq \varphi(wx_1 + (1 - w)x_2)$. Jensen's inequality generalizes this statement: if $X$ is a random variable and $\varphi$ is a concave function, then $\mathbb{E}[\varphi(X)] \leq \varphi(\mathbb{E}[X])$. The right panel illustrates the implications of Jensen's inequality for power, namely that $\mathbb{E}[P(\theta, n, \alpha)] \leq P(\mathbb{E}[\theta], n, \alpha)$ so that assessments of power based on the mean of $\theta$ are on average too high. In the plot, $\mathbb{E}$ appears as E because our plotting software lacks the facility to produce $\mathbb{E}$.

on all other values; consequently, $EP(\pi, n, \alpha) = P(\theta, n, \alpha)$. This is how the point estimate and safeguard power approaches function: they replace $\pi(\theta|D)$ with a probability mass of one at $\tilde{\theta}$ (i.e., the mean and twentieth percentile respectively). As discussed, this does not generally lead to the desired level of power on average. To obtain the desired level of power on average, it is instead necessary to choose $\tilde{\theta}$ such that $P(\tilde{\theta}, n^\star, \alpha) = EP^\star(\pi, n^\star, \alpha)$ where $EP^\star$ is the desired level of power on average and $n^\star$ is the sample size requisite to obtain it. We refer to $\tilde{\theta}$ chosen in this manner as the PCES and note, when the PCES is used in conjunction with textbook formulae or software, power is set to the desired level on average. Our proposed PCES is novel in that it is the unique effect size that accounts for the uncertainty about $\theta$ implied by $\pi(\theta|D)$ in a properly calibrated manner: sample sizes based on the PCES used in conjunction with textbook formulae or software provide the desired level of power on average.

We present an approach for analytically deriving the PCES in the appendix. Our approach is quite general and holds for a wide variety of cases. Using our analytic strategy, we derive PCES formulae for the statistical tests most often used in psychology (e.g., comparisons of independent and dependent means, comparisons of independent and dependent proportions, and tests of correlation coefficients) and present them in Table 2. To use these formulae, researchers need only a point estimate (also required by both the point estimate and safeguard power approaches) as well as an estimate of its error (also required by the safeguard power approach); thus, our approach requires no more information than comparable proposals in the literature while yielding better (i.e., calibrated) performance. Our formulae can be used to improve sample size planning as discussed above and as illustrated in the next section.

In addition to aiding in sample size planning, the PCES also provides intuition about the importance of uncertainty in a given example. For example, consider a researcher interested in a one-sided test of $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ at $\alpha = 0.05$ with 80% as the desired level of power (i.e., $\beta = 0.20$). In this case, the PCES is of the form $\tilde{\theta} = 2.05\theta - 1.05\sqrt{\theta^2 + 2.00\nu^2}$ (see the middle column of Table 2) where (a) $\theta$ denotes the point estimate of the parameter and (b) $\nu$ denotes the uncertainty in that point estimate.

We illustrate in Figure 2 the relationship between $\theta$, the point estimate, and $\tilde{\theta}$, the value to be used in conjunction with textbook formulae or software, for various values of $\nu$, the uncertainty in the point estimate; we do so for (a) the point estimate approach, (b) the safeguard power approach, and (c) the PCES approach. The plot shows that $\tilde{\theta}$ is equal to the point estimate $\theta$ for the point estimate approach; this is unsurprising since this approach ignores uncertainty $\nu$. In contrast, $\tilde{\theta}$ is equal to the point estimate $\theta$ plus a downward adjustment for both the safeguard power and PCES approaches. In both cases, the downward adjustment is larger when the uncertainty $\nu$ is larger; however, the downward adjustment for the safeguard power approach is constant given $\nu$ while the downward adjustment for the PCES is smaller the larger is the point estimate $\theta$. Thus, importantly, the PCES approach accounts for the size of both the point estimate $\theta$ and uncertainty $\nu$. In particular, the PCES $\tilde{\theta}$ is a roughly linear function of the point estimate $\theta$ when the point estimate is large relative to uncertainty $\nu$ (and indeed the PCES is $\theta$ for sufficiently large $\theta$); in contrast, when uncertainty $\nu$ is large relative to the point estimate $\theta$—the setting where accounting for uncertainty is most important for power—there is substantial nonlinearity, and the PCES $\tilde{\theta}$ can be quite a bit smaller than the point estimate $\theta$.

Figure 2 illustrates another important phenomenon, namely for $\nu$ sufficiently large relative to $\theta$ the PCES $\tilde{\theta}$ either does not exist or differs in sign with the point estimate. In such cases, the PCES is omitted from the plot, is not meaningful, and should not be used. Fortunately, these cases are easy to characterize. First, the PCES will not exist if the term under the square root sign is negative; this will not occur provided $\alpha < \beta$, so we recommend restricting to tests with $\alpha < \beta$. Second, the PCES will be of the wrong sign when uncertainty is sufficiently large relative to the point estimate that the desired level of expected power is simply not possible (e.g., $\tilde{\theta} = 2.05\theta - 1.05\sqrt{\theta^2+2.00\nu^2}$ will differ in sign from $\theta$ when $\theta \leq \frac{1.05}{2.05}\sqrt{\theta^2 + 2.00\nu^2}$); if we assume, as should almost always be the case in practice, that $\alpha < 0.50$ (i.e., so that Type I error is less than 50%) and $\beta < 0.50$ (i.e., so that power is greater than 50%), the PCES will be of the correct sign provided $\nu < \theta/|z_\beta|$. In sum, we recommend setting $\alpha < 0.50$, $\beta < 0.50$, and $\alpha < \beta$; in this case, the PCES will exist and be of the proper sign provided $\nu < \theta/|z_\beta|$, and thus our PCES approach can be used.

## Application to Choice Overload

In this section, we provide a tutorial on the sample size methodologies discussed above in the context of replication studies. We assume the researcher is interested in replicating the choice overload effect (i.e., that an increase in the number of options from which to choose can lead to adverse consequences such as a decrease in the likelihood of making a choice or the satisfaction with a choice; Iyengar & Lepper, 2000) and wishes to conduct a two-condition between-subjects study with equal sample size $n$ in each condition. We also assume the researcher wants to set the sample size to achieve 80% power for a comparison of two independent means using a one-sided test at $\alpha = 0.05$; in this case, the textbook sample size formula is $n = \frac{12.37\sigma^2}{\Delta^2}$ subjects per condition (see Hays, 1963 and Table 1) where $\Delta$ is the difference in the means between the conditions and $\sigma$ is the standard deviation of the individual-level observations, and the PCES formula is $\tilde{\Delta} = 2.05\Delta - 1.05\sqrt{\Delta^2 + 2.00\nu^2}$ (see Table 2) where $\nu$ is the uncertainty in $\Delta$. We discuss in turn (a) data collection, (b) setting sample sizes when one prior study is available, and (c) setting sample sizes when multiple prior studies are available.

We note that here and throughout we present two digits after the decimal point for noninteger real numbers and zero digits for integers unless otherwise noted. All calculations, however, are based on unrounded numbers. Consequently, calculations based on the rounded numbers presented in the text may not match exactly.

## Data Collection

In Table 3, we present data from three studies of the choice overload effect. The study design for each was a two-condition between-subjects study that measured satisfaction as a dependent variable. For the first study (Iyengar & Lepper, 2000; Study 2), the means and standard deviations in the table match those reported in the paper while the sample sizes in the table are the number of subjects that completed the assignment (i.e., the sample sizes

Table 2
*Power-Calibrated Effect Size Formulae for Common Tests*

| Test | General | One-sided ($\alpha = 0.05$, $\beta = 0.20$) | Two-sided ($\alpha = 0.05$, $\beta = 0.20$) |
|---|---|---|---|
| Comparison of two independent means | $\frac{z_{1-\alpha}\Delta + z_\beta\sqrt{\Delta^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}}{z_{1-\alpha} + z_\beta}$ | $2.05\Delta - 1.05\sqrt{\Delta^2 + 2.00\nu^2}$ | $1.75\Delta - 0.75\sqrt{\Delta^2 + 3.13\nu^2}$ |
| Comparison of two dependent means | $\frac{z_{1-\alpha}\Delta + z_\beta\sqrt{\Delta^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}}{z_{1-\alpha} + z_\beta}$ | $2.05\Delta - 1.05\sqrt{\Delta^2 + 2.00\nu^2}$ | $1.75\Delta - 0.75\sqrt{\Delta^2 + 3.13\nu^2}$ |
| Comparison of two independent proportions | $\frac{z_{1-\alpha}\Delta + z_\beta\sqrt{\Delta^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}}{z_{1-\alpha} + z_\beta}$ | $2.05\Delta - 1.05\sqrt{\Delta^2 + 2.00\nu^2}$ | $1.75\Delta - 0.75\sqrt{\Delta^2 + 3.13\nu^2}$ |
| Comparison of two dependent proportions | $\frac{1}{2} + \frac{z_{1-\alpha}\left(\ddot{p} - \frac{1}{2}\right) + z_\beta\sqrt{\left(\ddot{p} - \frac{1}{2}\right)^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}}{z_{1-\alpha} + z_\beta}$ | $\frac{1}{2} + 2.05\left(\ddot{p} - \frac{1}{2}\right) - 1.05\sqrt{\left(\ddot{p} - \frac{1}{2}\right)^2 + 2.00\nu^2}$ | $\frac{1}{2} + 1.75\left(\ddot{p} - \frac{1}{2}\right) - 0.75\sqrt{\left(\ddot{p} - \frac{1}{2}\right)^2 + 3.13\nu^2}$ |
| Test of a correlation coefficient | $\frac{z_{1-\alpha}Z_\rho + z_\beta\sqrt{Z_\rho^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}}{z_{1-\alpha} + z_\beta}$ | $2.05Z_\rho - 1.05\sqrt{Z_\rho^2 + 2.00\nu^2}$ | $1.75Z_\rho - 0.75\sqrt{Z_\rho^2 + 3.13\nu^2}$ |

*Note.* The null hypothesis is that of no difference or no correlation respectively. For two-sided tests, use $z_{1-\alpha/2}$ in place of $z_{1-\alpha}$ where $z_\gamma$ is the $100\gamma$ percentile of the standard normal distribution. In the table, $\Delta$ is the researcher's point estimate of the difference in means or proportions between the conditions, $\ddot{p}$ is the researcher's point estimate of the change proportion (i.e., $\ddot{p} = p_{10}/(p_{01} + p_{10})$ where $p_{ij}$ is the researcher's point estimate of the proportion who selected $i$ then $j$), $Z_\rho$ is the Fisher $Z$-transformation of the researcher's point estimate of the correlation coefficient (i.e., $Z_\rho = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) = \operatorname{arctanh}(\rho)$), and $\nu^2$ denotes the variance associated with the point estimate of $\Delta$, $\ddot{p}$, and $Z_\rho$ respectively. Our formulae assume without loss of generality that $\Delta$ and $\left(\ddot{p} - \frac{1}{2}\right)$ are positive; if $Z_\rho$ is negative use $|Z_\rho|$ in place of it.

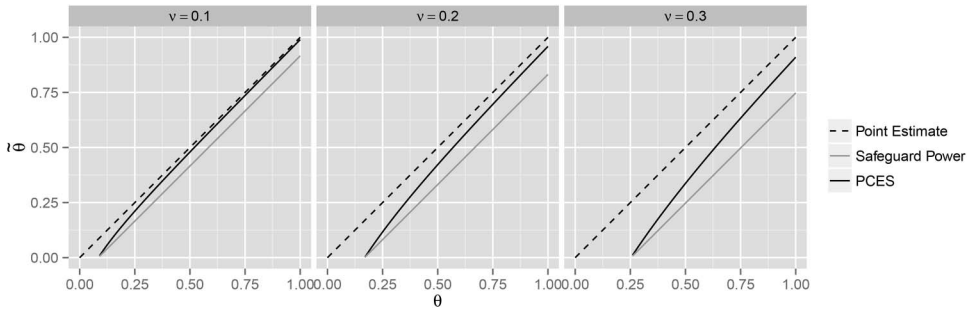



*Figure 2.* $\tilde{\theta}$ versus $\theta$. We assume a one-sided test at $\alpha = 0.05$ with 80% as the desired level of power.



reported in the paper multiplied by the completion rate reported in the paper). For the second study (Fasolo, Carmeci, & Misuraca, 2009; Study 1), the means, standard deviations, and sample sizes match those reported in the paper. For the third study (Diehl & Poynor, 2010; Study 2), the means and sample sizes in the table match those reported in the paper while the standard deviations in the table were determined from the means and sample sizes as well as the $F$-statistic of 4.18 reported in the paper $\left(\frac{7.81-7.40}{\sqrt{4.18(1/78+1/87)}} = 1.29\right)$. The first study measured satisfaction on a 10-point scale, the second on a 9-point scale, and the third on a 5-point scale.

We chose these studies because they were highly suitable for illustration purposes: (a) they reported the data cleanly and clearly, (b) they followed the same study design, (c) they measured the same dependent variable, and (d) they used a very similar measurement scale for the dependent variable. We also note that the only distinction relevant for the sample size planning is whether there is one prior study or multiple (i.e., more than one) prior studies in the domain; as shown below, the inputs for the sample size and PCES formulae are based on the data from the single study in the former case and a meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper & Hedges, 1994; Cooper, Hedges, & Valentine, 2009; Hedges & Olkin, 1985; Hunter & Schmidt, 1990) of the data from the set of studies in the latter case.

## One Prior Study

We show how to use the sample size and PCES formulae in the context of one prior study, which we take to be the first study (Iyengar & Lepper, 2000; Study 2). The formulae require three inputs in total (i.e., $\Delta$ and $\sigma$ for the sample size formula and $\Delta$ and $\nu$ for the PCES formula), and we show how to obtain them from the data.

With a single study, $\Delta$ can be set to the observed difference in the means ($8.09 - 7.69 = 0.40$) while $\sigma$ can be set to the pooled standard deviation $\left(\sqrt{\frac{(52-1)1.05^2+(74-1)0.82^2}{52+74-2}} = 0.92\right)$. The uncertainty in the difference in the means $\nu$ can be set to the standard error of the observed difference in the means $\left(\sqrt{\frac{0.92^2}{52}+\frac{0.92^2}{74}} = 0.17\right)$. We note that all three of these values can be easily obtained with standard statistical software.

The point estimate approach ignores uncertainty in $\Delta$ and uses the point estimate of $\Delta$ in conjunction with the sample size formula $\frac{12.37\sigma^2}{\Delta^2}$. Thus, the sample size suggested by the point estimate approach is $\frac{12.37 \cdot 0.92^2}{0.40^2} = 66$ subjects per condition.

The safeguard power approach accounts for uncertainty in $\Delta$ by using the twentieth percentile $\Delta + z_{0.20}\nu = 0.40 - 0.84 \cdot 0.17 = 0.26$ in conjunction with the sample size formula. Thus, the sample size suggested by the safeguard power approach is $\frac{12.37 \cdot 0.92^2}{0.26^2} = 156$ subjects per condition.

Finally, the PCES approach accounts for uncertainty in a manner that provides the desired level of power on average. The PCES is $2.05\Delta - 1.05\sqrt{\Delta^2 + 2.00\nu^2} = 2.05 \cdot 0.40 - 1.05\sqrt{0.40^2 + 2.00 \cdot 0.17^2} = 0.33$. Thus, the sample size suggested by the PCES approach is $\frac{12.37 \cdot 0.92^2}{0.33^2} = 95$ subjects per condition. The PCES approach yields the desired level of power on average and strikes a balance between the point estimate approach that uses too few subjects and thus has too little power on average and the safeguard power approach that uses too many subjects and thus has too much power on average.

We note that here and below we first determined the appropriate effect size (i.e., 0.40 for the point estimate approach, 0.26 for the safeguard power approach, and 0.33 for the PCES approach) and then used it in conjunction with the sample size formula. An alternative approach for those who use statistical software for

Table 3
*Choice Overload Data*

| Article | Study | Small choice set | | | Large choice set | | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | $n$ | Mean | SD | $n$ |
| Iyengar and Lepper (2000) | Study 2 | 8.09 | 1.05 | 52 | 7.69 | 0.82 | 74 |
| Fasolo et al. (2009) | Study 1 | 3.81 | 0.54 | 32 | 3.78 | 0.55 | 32 |
| Diehl and Poynor (2010) | Study 2 | 7.81 | 1.29 | 78 | 7.40 | 1.29 | 87 |

*Note.* The study column indicates the study number in the article.

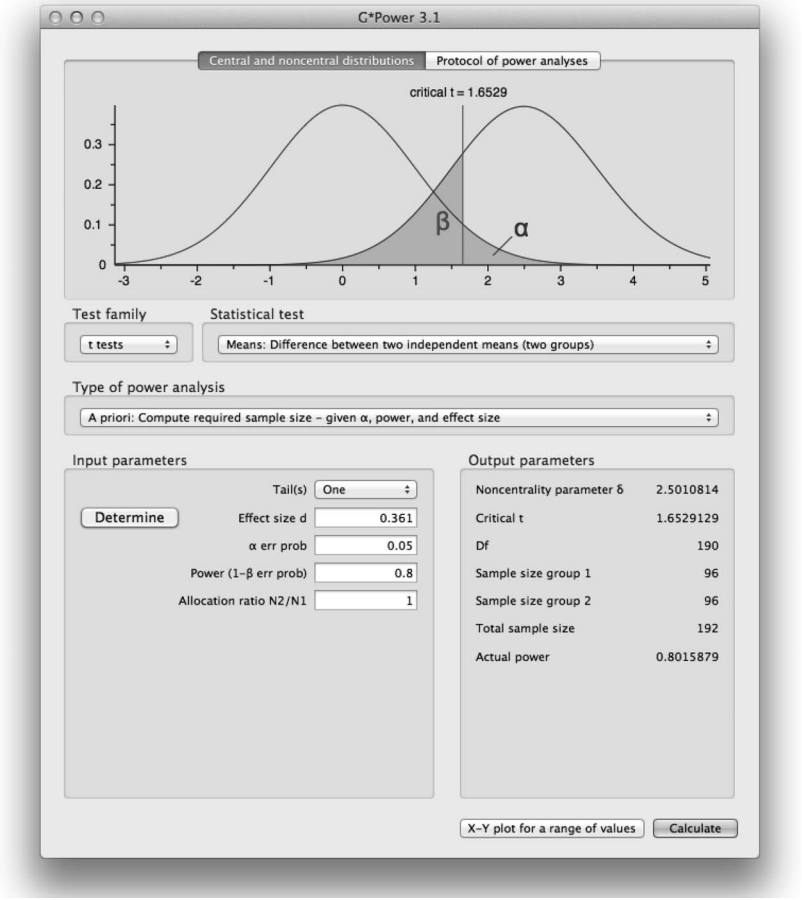



*Figure 3.* Screenshot of G*Power using the standardized power-calibrated effect size.

sample size planning would be to first determine the appropriate effect size and then use it with the relevant software instead of the formula; we provide an example of how to do this for the PCES approach using G*Power in Figure 3, noting that G*Power requires the use of the standardized (rather than unstandardized) effect size $\Delta/\sigma = 0.33/0.92 = 0.3610$ (we present four digits for greater precision). A third possibility that we discuss in detail below is to use our website or code.

## Multiple Prior Studies

We show how to use the sample size and PCES formulae in the context of multiple prior studies, which we take to be the three studies reported in Table 3. The formulae require the same three inputs as in the case of one prior study, and we show how to obtain them from a meta-analysis of the data.

To obtain the three inputs, we conducted a meta-analysis using the metafor package (Viechtbauer, 2010) in R (for code to replicate this meta-analysis see the online supplementary materials or our website). The meta-analysis yielded an effect size estimate of 0.31 with a standard error of 0.11, which we can use for $\Delta$ and $\nu$ respectively.[1] As the meta-analysis was conducted on the standardized mean difference scale, $\sigma$ can be set to one.

The point estimate approach thus suggests a sample size of $\frac{12.37 \cdot 1.00^2}{0.31^2} = 131$ subjects per condition. The safeguard power approach uses $0.31 - 0.84 \cdot 0.11 = 0.22$ as the effect size and thus suggests a sample size of $\frac{12.37 \cdot 1.00^2}{0.22^2} = 262$ subjects per condition. Finally, the PCES is $2.05 \cdot 0.31 - 1.05\sqrt{0.31^2 + 2.00 \cdot 0.11^2} = 0.27$, and thus the PCES approach suggests a sample size of $\frac{12.37 \cdot 1.00^2}{0.27^2} = 169$ subjects per condition. Again, the PCES approach yields the desired level of power on average and strikes a balance between the point estimate approach that uses too few subjects and thus has too little power on average and the safeguard power approach that uses too many subjects and thus has too much power on average.

[1] A random effects meta-analysis estimated by metafor using restricted (or residual or reduced) maximum likelihood (REML; Harville, 1977) estimated heterogeneity at zero, so we used a fixed effects meta-analysis. Given the paucity of studies in the meta-analysis, an estimate of zero heterogeneity is not necessarily surprising, and Bayesian approaches that bound the estimate away from zero can prove helpful (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013; Gelman, 2006). To account for heterogeneity (McShane & Böckenholt, 2014) as well as uncertainty in sample size planning, researchers can use as uncertainty $\nu$ the square root of the sum of (a) heterogeneity (represented as a variance) and (b) the variance of the effect size estimate (i.e., the square of the standard error).

We note that the sample size requisite to obtain the desired level of power based on the meta-analysis is larger than that based on the first study for all three approaches. This occurs because, while the meta-analysis reduced uncertainty ($\nu/\sigma = 0.18$ for the first study versus $\nu/\sigma = 0.11$ for the meta-analysis when the point estimate is presented on the standardized scale), it also reduced the point estimate ($\Delta/\sigma = 0.36$ for the first study versus $\Delta/\sigma = 0.31$ for the meta-analysis when the point estimate is presented on the standardized scale). This latter effect dominates in this example thereby leading to the larger sample size.

## Exploring the Effect of Uncertainty

We explore how the sample size requisite for 80% power and the expected power vary for different values of uncertainty $\nu$ in Figure 4 assuming the effect size is centered on the value given by the meta-analysis of the choice overload data discussed above; the points in the figure denote the value of $\nu$ obtained from the meta-analysis. As can be seen, when $\nu$ is near zero (i.e., the researcher is relatively certain in the effect size $\Delta$), all three approaches converge: they call for the same sample size and provide the desired level of power. As $\nu$ grows, however, each of the three approaches behaves quite differently.

The point estimate approach calls for the same sample size regardless of the value of $\nu$. This is because this approach does not account for uncertainty. As a consequence, the expected power drops below the desired level when $\nu$ is greater than zero.

On the other hand, the safeguard power approach calls for a sample size that grows along with $\nu$: more subjects are requisite when there is greater uncertainty about the underlying effect size. However, because this approach rather pessimistically assumes an effect size that is at the twentieth percentile, the sample size grows quite sharply and might be considered untenable even for relatively modest values of $\nu$. Further, because of the large sample size, the expected power is above the desired level when $\nu$ is greater than zero. We also observe that the expected power is non-monotone in $\nu$; this is due to the difference in variance between the null and alternative hypotheses in the expected power setup (see Appendix A and, in particular, the discussion surrounding Equation 4).

Finally, our PCES approach, which accounts for uncertainty in a manner that yields the desired level of power on average, appears to strike a nice balance between the other two approaches. Like the safeguard power approach, the PCES approach requires a larger sample size when uncertainty is larger. However, the penalty is not nearly as dramatic, and the sample size remains reasonable even for relatively large values of $\nu$. Further, expected power is properly calibrated at the desired level.

# Facilitating the PCES

In this section, we discuss an easy-to-use website that implements the PCES approach to sample size planning, and we reproduce the choice overload results in the context of this website. We also discuss the principal code underlying the website. Finally, we provide several additional examples.

## Website and Code

To facilitate sample size planning in the face of uncertainty in a manner that properly calibrates power, we have created an easy-to-use website that implements our PCES and sample size formulae for the cases outlined in Table 2. The website is available at https://blakemcshane.shinyapps.io/pces/, and it contains a tutorial that explains how to reproduce the choice overload results contained in this paper. By following the tutorial as well as the additional instructional material on the website, researchers should easily be able to account for uncertainty in their own sample size determinations.

To reproduce the choice overload results in the context of the website, a researcher must first select the parameters for the sample size analysis; in this case, the researcher wants to (a) compare two independent means using (b) a one-sided test (c) at $\alpha = 0.05$ (d) with 80% as the desired level of power (all four of

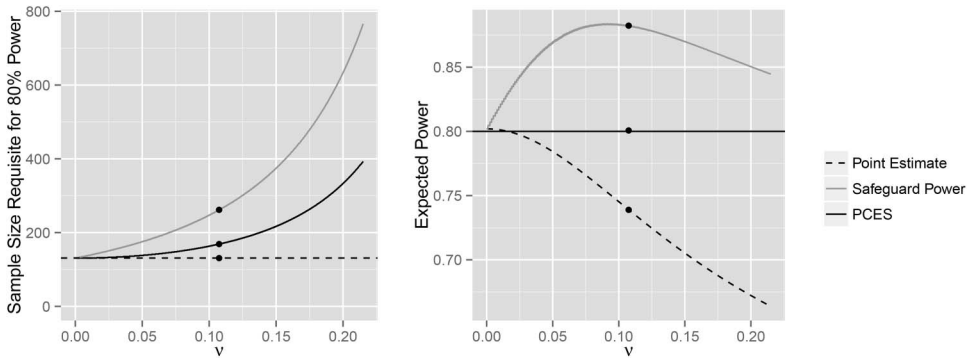



*Figure 4.* Sample size per condition requisite for 80% power and expected power. We assume a comparison of two independent means using a one-sided test at $\alpha = 0.05$ with 80% as the desired level of power and assume the effect size is centered around $\Delta = 0.31$ as per the meta-analysis of the choice overload data with the uncertainty $\nu$ given on the $x$-axis. We also assume $\sigma = 1$ as per the meta-analysis. The points represent the sample size per condition requisite for 80% power and the expected power for each approach at the $\nu$ obtained from the meta-analysis (i.e., $\nu = 0.11$). The various approaches lead to quite different sample sizes even for relatively modest values of uncertainty $\nu$. The expected power for the point estimate and safeguard power approaches can deviate substantially from the desired level of 80%.

these parameters can be adjusted on the website). Next, the researcher provides the three inputs $\Delta$, $\sigma^2$, and $\nu^2$, which are 0.4000, $0.92^2 = 0.8493$, and $0.17^2 = 0.0278$ respectively when there is one prior study and 0.3081, 1.0000, and $0.11^2 = 0.0116$ respectively when there are multiple prior studies (we present four digits to facilitate exact replication of the results reported in the prior section). Finally, the website returns the output displayed in Figure 5. The first column returns the effect size and sample size that do not account for uncertainty (i.e., the point estimate approach) while the second column returns the PCES and sample size that account for uncertainty.

To further facilitate the PCES approach, we have provided both in the online supplementary materials and on the website the principal code that underlies the website as well as code to replicate the choice overload results. The code provides a detailed set of variable definitions. It also implements the textbook sample size formulae reported in Table 1, the PCES formulae reported in Table 2, and power-calibrated sample size formulae. It further shows how to replicate the website defaults thereby illustrating how to use the code. Finally, it shows how to replicate the choice overload results.

We believe these resources will allow researchers to easily account for uncertainty in their own sample size determinations.

## Additional Examples

The choice overload example was an in-depth example of a comparison of two independent means resulting from a between-subjects study with a continuous dependent variable. In this section, we provide brief examples for the four other cases considered in Tables 1–2 again assuming a one-sided test at $\alpha = 0.05$ with 80% as the desired level of power.

**Comparison of two dependent means.** A comparison of two dependent means resulting from a within-subjects study with a continuous dependent variable is very similar to a comparison of two independent means. The textbook sample size formula is $\frac{6.18\sigma_D^2}{\Delta^2}$ where $\sigma_D$ is the standard deviation of the individual-level difference scores, and the PCES formula is again $2.05\Delta - 1.05\sqrt{\Delta^2 + 2.00\nu^2}$. Assuming a prior study (or a meta-analysis of several prior studies) yields a point estimate $\Delta = 0.20$, standard deviation $\sigma_D = 1$, and standard error (of $\Delta$) $\nu = 0.10$, the PCES is $2.05 \cdot 0.20 - 1.05\sqrt{0.20^2 + 2.00 \cdot 0.10^2} = 0.15$, and the PCES approach suggests a sample size of $\frac{6.18 \cdot 1.00^2}{0.15^2} = 265$ subjects in total. We note that, for those seeking to obtain these three inputs, meta-analysis can accommodate a set of prior studies that includes both between-subjects and within-subjects designs (Gibbons, Hedeker, & David, 1993).

**Comparison of two independent proportions.** A comparison of two independent proportions resulting from a between-subjects study with a binary dependent variable also requires three inputs: $p_1$ and $p_2$ (where $p_i$ is the proportion in condition $i$) as well as $\nu$. The textbook sample size formula is $\frac{12.37\bar{p}(1-\bar{p})}{\Delta^2}$ where $\bar{p} = (p_1 + p_2)/2$ and the PCES formula is again $2.05\Delta - 1.05\sqrt{\Delta^2 + 2.00\nu^2}$ where $\Delta = p_2 - p_1$. Assuming a prior study (or a meta-analysis of several prior studies) yields a point estimate $p_1 = 0.40$ and $p_2 = 0.60$ (so that $\bar{p} = 0.50$ and $\Delta = 0.20$) and standard error (of $\Delta$) $\nu = 0.10$, the PCES is $2.05 \cdot 0.20 -$

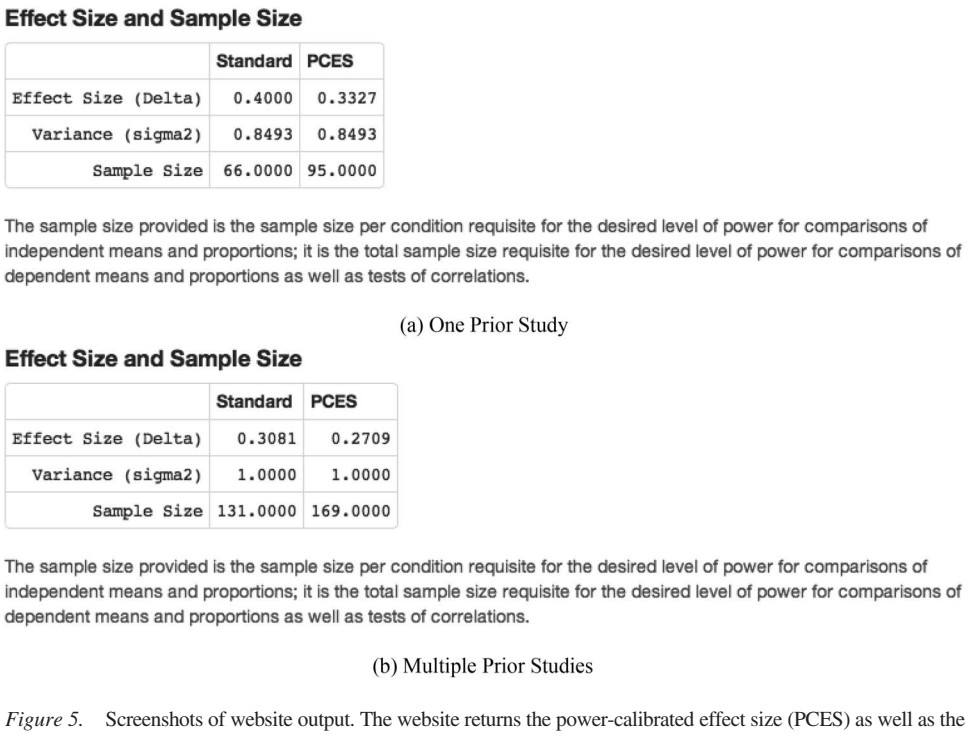



*Figure 5.* Screenshots of website output. The website returns the power-calibrated effect size (PCES) as well as the sample size required for adequate power. This analysis assumes a comparison two independent means using a one-sided test at $\alpha = 0.05$ with 80% as the desired level of power based on one prior study and multiple prior studies respectively.

$1.05\sqrt{0.20^2 + 2.00 \cdot 0.10^2} = 0.15$, and the PCES approach suggests a sample size of $\frac{12.37 \cdot 0.50 \cdot (1 - 0.50)}{0.15^2} = 133$ subjects per condition.

**Comparison of two dependent proportions.** A comparison of two dependent proportions resulting from a within-subjects study with a binary dependent variable yet again requires three inputs: $p_{01}$ and $p_{10}$ (where $p_{ij}$ is the proportion who selected $i$ then $j$) as well as $v$. The textbook sample size formula is $\frac{1.55}{(\ddot{p} - \frac{1}{2})^2(p_{01} + p_{10})}$ where $\ddot{p} = p_{10}/(p_{01} + p_{10})$ and the PCES formula is $\frac{1}{2} + 2.05\left(\ddot{p} - \frac{1}{2}\right) - 1.05\sqrt{\left(\ddot{p} - \frac{1}{2}\right)^2 + 2.00v^2}$. Assuming a prior study (or a meta-analysis of several prior studies) yields a point estimate $p_{01} = 0.10$ and $p_{10} = 0.20$ (so that $\ddot{p} = 0.67$) and standard error (of $\ddot{p}$) $v = 0.10$, the PCES is $\frac{1}{2} + 2.05 \cdot \left(0.67 - \frac{1}{2}\right) - 1.05\sqrt{\left(0.67 - \frac{1}{2}\right)^2 + 2.00 \cdot 0.10^2} = 0.61$, and the PCES approach suggests a sample size of $\frac{1.55}{\left(0.61 - \frac{1}{2}\right)^2(0.10 + 0.30)} = 409$ subjects in total.

**Test of a correlation coefficient.** Finally, a test of a correlation coefficient resulting from a study that collects two continuous dependent variables per subject requires two inputs: $\rho$ and $v$. The textbook sample size formula is $\frac{6.18}{Z_\rho^2} + 3$ where $Z_\rho$ is the Fisher $Z$-transformation of the correlation coefficient (i.e., $Z_\rho = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) = \operatorname{arctanh}(\rho)$) and the PCES formula is $2.05Z_\rho - 1.05\sqrt{Z_\rho^2 + 2.00v^2}$. Assuming a prior study (or a meta-analysis of several prior studies) yields a point estimate $\rho = 0.20$ (and thus $Z_\rho = 0.20$) and standard error (of $Z_\rho$) $v = 0.10$, the PCES is $2.05 \cdot 0.20 - 1.05\sqrt{0.20^2 + 2.00 \cdot 0.10^2} = 0.16$, and the PCES approach suggests a sample size of $\frac{6.18}{0.16^2} + 3 = 257$ subjects in total.

**Inputs.** We note that when only one prior study is available, the inputs for the four cases discussed above can easily be obtained from that study just as illustrated for the case of a comparison of two independent means using data from Iyengar and Lepper (2000) Study 2. For example, for a test of a correlation coefficient, the correlation observed in the prior study yields the point estimate of $\rho$ (and thus of $Z_\rho$) while the uncertainty is given by the standard error $v = \sqrt{1/(n_0 - 3)}$ where $n_0$ is the sample size from that study. When multiple prior studies are available, a meta-analysis of those studies yields the necessary inputs.

## Discussion

Researchers planning replication studies want to ensure their studies have adequate power. Textbook sample size formulae and statistical software guarantee the desired level of power but presume the effect size is known. Although researchers often have some idea of the effect size from prior studies, they still face considerable uncertainty. We demonstrate how to cope with and explicitly account for this uncertainty in a manner that properly calibrates power via our PCES approach. As a result, this paper contributes to previous work on effect sizes that stress the importance of accompanying an effect size with an interval estimate as a measure of estimation uncertainty (Kelley & Preacher, 2012) as well work that discusses how to set sample sizes in the face of uncertain effect sizes (Perugini et al., 2014).

To make our approach directly applicable to applied research, we provided PCES formulae for the statistical tests most often used in psychology in Table 2. The PCES can be used in conjunction with textbook formulae or software to obtain sample sizes that account for uncertainty in a manner that properly calibrates power. Our approach requires larger sample sizes than the point estimate approach that ignores uncertainty, but it does not require the more extreme sample sizes associated with alternative approaches that account for uncertainty such as the safeguard power approach. We also showed how to obtain the inputs to our formulae in the context of planning a replication study of the choice overload effect when one prior study is available and when multiple prior studies are available. Finally, we discussed an easy-to-use website and code that implements our formulae.

When researchers assume a known effect size but actually face uncertainty, power calculations are optimistic, and this false optimism causes sample sizes to be set too low. Consequently, studies may fail at a greater than expected rate, thus shedding light on current difficulties in replicating psychological research. The formulae given in Tables 1–2 and implemented on the website and in the code provided both in the online supplementary materials and on the website allow researchers to adjust an effect size for uncertainty and then to use textbook formulae or software. This restores power to the desired level on average and thus should be helpful for researchers attempting replication.

## References


Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, United Kingdom: Wiley.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217–224.

Brodeur, A., Le, M., Sangnier, M., & Zylberberg, Y. (2012). *Star wars: The empirics strike back*. Paris, France: Paris School of Economics.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A non-degenerate estimator for hierarchical variance parameters via penalized likelihood estimation. *Psychometrika, 78,* 685–709.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29.

Diehl, K., & Poynor, C. (2010). Great expectations?! Assortment size, expectations and satisfaction. *Journal of Marketing Research, 47,* 312–322.

Fasolo, B., Carmeci, F. A., & Misuraca, R. (2009). The effect of choice complexity on perception of time spent choosing: When choice takes longer but feels shorter. *Psychology and Marketing, 26,* 213–228.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57,* 153–169.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1,* 515–533.

Gibbons, R. D., Hedeker, D. R., & David, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics, 18,* 271–279.

Gillett, R. (1994). An average power criterion for sample size estimation. *Journal of the Royal Statistical Society, Series D (The Statistician), 43,* 389–394.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72,* 320–338.

Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehart, and Winston.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79,* 996–1006.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica, 30,* 175–193.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science, 9,* 612–625.

O'Hagan, A., Stevens, J. W., & Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics, 4,* 187–201.

Pashler, H., & Wagenmakers, E.-J. (2012). Eds.' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9,* 319–332.

R Core Team. (2012). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, United Kingdom: Wiley.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Yong, E. (2012). Replication studies: Bad copy. *Nature, 485,* 298–300.

(*Appendix follows*)

## Appendix

## Expected Power and the PCES

As discussed in the main text, the PCES is the $\tilde{\theta}$ such that $P(\tilde{\theta}, n^\star, \alpha) = EP^\star(\pi, n^\star, \alpha)$ where $EP^\star$ is the desired level of power on average and $n^\star$ is the sample size requisite to obtain it. To determine the PCES, we return to the expected power equation (Equation 2) noting

$$\begin{aligned} EP(\pi, n, \alpha) &= \int_\Theta P(\theta, n, \alpha)\pi(\theta \mid D)d\theta \\ &= \int_\Theta \left[\int_{\mathbf{x}:T(\mathbf{x})\in R_{\alpha,n}} f_T(\mathbf{x} \mid \theta, n)d\mathbf{x}\right]\pi(\theta \mid D)d\theta \\ &= \int_{\mathbf{x}:T(\mathbf{x})\in R_{\alpha,n}} \int_\Theta f_T(\mathbf{x} \mid \theta, n)\pi(\theta \mid D)d\theta d\mathbf{x} \\ &= \int_{\mathbf{x}:T(\mathbf{x})\in R_{\alpha,n}} \left[\int_\Theta f_T(\mathbf{x} \mid \theta, n)\pi(\theta \mid D)d\theta\right]d\mathbf{x} \\ &= \int_{\mathbf{x}:T(\mathbf{x})\in R_{\alpha,n}} g_T(\mathbf{x} \mid \pi, n, D)d\mathbf{x} \end{aligned} \quad (3)$$

where the first line defines expected power as a weighted average of power, the second line simply substitutes in the definition of power (Equation 1), the third and fourth lines rearrange terms, and the fifth line defines $g_T(\mathbf{x} \mid \pi, n, D) = \int_\Theta f_T(\mathbf{x} \mid \theta, n)\pi(\theta \mid D)d\theta$; $g_T(\mathbf{x} \mid \pi, n, D)$ is sometimes called the posterior predictive distribution of the test statistic.

Equation 3 contains two principal probability density functions: (a) $f_T$, the density of the sampling distribution of the test statistic, and (b) $\pi$, the density that characterizes the researcher's beliefs about the parameter $\theta$ (sometimes $\pi$ is called the prior distribution; as illustrated in the main text, this is misleading because $\pi$ is typically based on data from prior studies and thus is better thought of as the posterior distribution resulting from the observation of those studies). As shown in the equation, $f_T$ and $\pi$ jointly determine $g_T$.

When $f_T$ and $\pi$ are both normal densities and $\theta$ denotes the mean of $f_T$, $g_T$ is also a normal density (with mean equal to the mean of $\pi$ and variance equal to the sum of the variances of $f_T$ and $\pi$), and EP can typically be simplified considerably (e.g., to the integral of a scalar variable from a given value to infinity). While assuming both $f_T$ and $\pi$ are normal may seem like a large assumption, in practice this is not necessarily the case. First, $f_T$ is the density of the sampling distribution of a test statistic, and such sampling distributions are often normal; when they are not, they are often asymptotically normal. Second, estimates of $\theta$ are often derived from prior research and in particular from statistics that have normal or asymptotically normal sampling distributions (in fact in some cases these distributions can even be used as $\pi$). Consequently, it is reasonable across a wide variety of cases to assume that both $f_T$ and $\pi$ are normal. Further, as we demonstrate below, the assumption works well in practice.

Now, consider a one-sided test of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ where (a) $f_T$ is normal with variance $\sigma_0^2(n)$ that is a function of the sample size $n$ and (b) $\pi$ is normal with mean $\theta_1$ (reflecting the point estimate of $\theta$) and variance $\nu^2$ (reflecting the uncertainty in the point estimate). Let $\sigma_1^2(n) = \sigma_0^2(n) + \nu^2$. To achieve a given level of power $1 - \beta$ while maintaining the size $\alpha$ of the test, we simply need to choose the sample size $n$ such that

$$\theta_0 + z_{1-\alpha}\sigma_0(n) = \theta_1 + z_\beta\sigma_1(n) \quad (4)$$

where $z_\gamma$ is the $100\gamma$ percentile of the standard normal distribution. We note that (a) these tests assume $\theta_1 > \theta_0$ without loss of generality and (b) this approach also applies to two-sided tests.[2]

When $\theta$ is known to be $\theta_1$, then $\nu^2 = 0$ so $\sigma_1^2(n) = \sigma_0^2(n)$. Consequently, we can solve Equation 4 for $\sigma_0(n) = \frac{\theta_1 - \theta_0}{z_{1-\alpha} - z_\beta}$. For normal $f_T$, we typically have $\sigma_0(n) = c/\sqrt{n}$ for some known constant $c$ thereby allowing us to solve for $n$ yielding

$$n = \frac{c^2(z_{1-\alpha} - z_\beta)^2}{(\theta_1 - \theta_0)^2}. \quad (5)$$

Thus, when $\theta$ is known to be $\theta_1$, we can achieve a given level of power by setting sample sizes according to the above formula.

[2] For two-sided tests of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, we can use the approach outlined above with $z_{1-\alpha/2}$ in place of $z_{1-\alpha}$. While this excludes the generally small amount of probability that lies in the lower tail (i.e., below $\theta_0 + z_{\alpha/2}\sigma_0(n)$), we believe this is reasonable in our setting of replication studies as a replication study that attained statistical significance but was opposite in sign to a prior study would not be deemed a successful replication.

(*Appendix continues*)

When $\theta$ is unknown but believed to have expectation $\theta_1$ and variance $\nu^2$ (i.e., $\pi$ has mean $\theta_1$ and variance $\nu^2$), then $\sigma_1^2(n) = \sigma_0^2(n) + \nu^2$. Substituting the square root of this quantity into Equation 4 for $\sigma_1(n)$ and solving for $\sigma_0(n)$ yields

$$\sigma_0(n) = \frac{z_{1-\alpha}(\theta_1 - \theta_0) + z_\beta\sqrt{(\theta_1 - \theta_0)^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}}{z_{1-\alpha}^2 - z_\beta^2}. \tag{6}$$

Again letting $\sigma_0(n) = c/\sqrt{n}$ for some known constant $c$, we can solve for $n$ yielding

$$n = \frac{c^2(z_{1-\alpha}^2 - z_\beta^2)^2}{\left[z_{1-\alpha}(\theta_1 - \theta_0) + z_\beta\sqrt{(\theta_1 - \theta_0)^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}\,\right]^2}. \tag{7}$$

Thus, when $\theta$ is unknown but believed to have expectation $\theta_1$ and variance $\nu^2$, we can achieve a given level of power by setting sample sizes according to the above formula.

To obtain the PCES, denoted $\tilde{\theta}_1$, we merely need to set Equation 5 (with $\tilde{\theta}_1$ in place of $\theta_1$) equal to Equation 7

$$\frac{c^2(z_{1-\alpha} - z_\beta)^2}{(\tilde{\theta}_1 - \theta_0)^2} = \frac{c^2(z_{1-\alpha}^2 - z_\beta^2)^2}{\left[z_{1-\alpha}(\theta_1 - \theta_0) + z_\beta\sqrt{(\theta_1 - \theta_0)^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}\,\right]^2} \tag{8}$$

and solve for $\tilde{\theta}_1$ yielding

$$\tilde{\theta}_1 = \theta_0 + \frac{z_{1-\alpha}(\theta_1 - \theta_0) + z_\beta\sqrt{(\theta_1 - \theta_0)^2 + \nu^2(z_{1-\alpha}^2 - z_\beta^2)}}{z_{1-\alpha} + z_\beta}. \tag{9}$$

This formula is easy to apply in practice. For instance, in the standard case of a null hypothesis of zero effect (i.e., $\theta_0 = 0$) with $\alpha = 0.05$ (i.e., $z_{1-\alpha} = 1.64$) and power equal to 80% (i.e., $\beta = 0.20$ and $z_\beta = -0.84$), it simplifies to

$$\tilde{\theta}_1 = 2.05\theta_1 - 1.05\sqrt{\theta_1^2 + 2.00\nu^2}. \tag{10}$$

The PCES $\tilde{\theta}_1$ used in conjunction with textbook sample size formulae (of the form of Equation 5) or statistical software yields sample sizes that account for uncertainty in $\theta_1$ and provide the desired level on power on average. The PCES also provides intuition about importance of uncertainty in a given example (see, for example, Figure 2).

The PCES formulae reported in Table 2 come directly from Equation 9 substituting as appropriate. In particular, for a comparison of two independent means, a comparison of two dependent means, and a comparison of two independent proportions, we substitute $\theta_0 = 0$ and $\theta_1 = \Delta$ (i.e., the difference between the two means or proportions). For a comparison of two dependent proportions, we substitute $\theta_0 = 1/2$ and $\theta_1 = \ddot{p}$. Finally, for a test of a correlation coefficient, we substitute $\theta_0 = 0$ and $\theta_1 = Z_\rho$.

A potential concern with the PCES formulae reported in Table 2 is that they are based on the assumption that both $f_T$ and $\pi$ are normal, which seldom holds precisely in practice. Consequently, sample sizes set based on these formulae may fail to provide the desired level of power on average. We can investigate this by noting that the expected power of a replication study based on a prior study with a sample size of $n_0$ is

$$\mathrm{EP}(n_0, n^\star, \alpha) = \int_{\mathcal{X}_0}\int_{\mathbf{x}:T(\mathbf{x})\in R_{\alpha,n}} f_T(\mathbf{x}\,|\,\theta, n^\star(\mathbf{x}_0))f_T(\mathbf{x}_0\,|\,\theta, n_0)d\mathbf{x}d\mathbf{x}_0 \tag{11}$$

where $\mathcal{X}_0$ is the sample space of the prior study data and $n^\star(\mathbf{x}_0)$ is the sample size for the replication study; $n^\star(\mathbf{x}_0)$ is generally a function of the observed prior study data $\mathbf{x}_0$.

For example, consider a comparison of two independent means. Assuming the individual-level observations are normally distributed with unknown variance (i.e., the standard assumption used in practice), then $f_T$ is not normal as assumed by our PCES formulae but noncentral $t$ with $2n - 2$ degrees of freedom and noncentrality parameter $\frac{\Delta}{\sqrt{2\sigma^2/n}}$ where $\Delta$ is the true difference in the two means, $\sigma^2$ is the variance of the individual-level observations, and $n$ is the sample size in each condition. To assess the performance of our formulae in this setting, we can calculate the expected power using Equation 11 with the noncentral $t$ distribution in place of $f_T$ and with the sample size based on the PCES for a comparison of two independent means in place of $n^\star(\mathbf{x}_0)$.

We depict this assessment in Figure A1 assuming a one-sided test at $\alpha = 0.05$ with 80% as the desired level of power; we set $\Delta$ to 0.2, 0.5, and 0.8 in turn and without loss of generality $\sigma^2$ to one so that the $\Delta$ correspond to the respective definitions of small, medium, and large effect sizes in psychology (Cohen, 1992). We note that $\nu^2 = 2\sigma^2/n$ so the uncertainty $\nu$ presented on the $x$-axis in Figure A1 corresponds to a prior study with sample size per condition ranging from infinite ($\nu = 0.0$) to about 22 ($\nu = 0.3$) while the expected power presented on the $y$-axis is that of the replication study when the sample size is set based on the PCES, which is in turn based on the data from the prior study. As can be seen, power is closely calibrated to the desired level: when $\nu$ is small or modest relative to $\Delta$, the sample size based on the PCES provides the desired level of power on average while when $\nu$ is large relative to $\Delta$, it provides somewhat greater power suggesting it is conservative in these settings.

We have conducted this assessment for all five PCES formulae reported in Table 2 across a wide range of values for $\alpha$, $\beta$, effect sizes (i.e., $\Delta$, $\ddot{p}$, and $Z_\rho$), and $\nu$, and confirm that the results presented in Figure A1 are representative.

(*Appendix continues*)

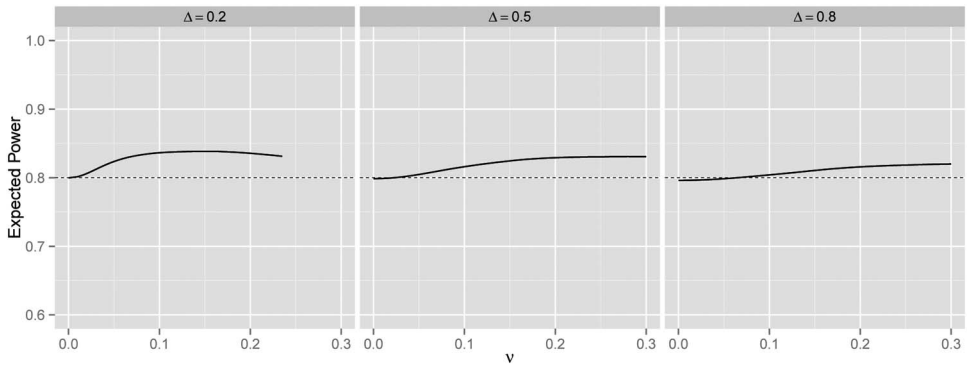



*Figure A1.* Expected power. We assume a comparison of two independent means using a one-sided test at $\alpha = 0.05$ with 80% as the desired level of power. Power is closely calibrated to the desired level. When $\nu$ is sufficiently large relative to $\Delta$, the power-calibrated effect size does not exist and so expected power is not defined.

As a final comment, recall that the PCES formulae reported in Table 2 are based on $\pi(\theta|D)$ being normal with mean $\theta$ and variance $\nu^2$. When $\theta$ is set to the point estimate based on data from prior studies and $\nu^2$ is set to the variance associated with that point estimate (e.g., as in the choice overload example in the main text), then $\pi(\theta|D)$ is the posterior distribution corresponding to a normal sampling distribution for the point estimate and a noninformative prior for $\theta$. In practice, a normal sampling distribution and a noninformative prior distribution need not be assumed, and PCESs can be derived under alternative specifications (although analytic PCES formulae will not always result). Moreover, if a normal sampling distribution and a normal prior are assumed, then our PCES formulae are still valid but simply require different inputs. In particular, if the prior is assumed to be normal with mean $\bar{\theta}$ and variance $\bar{\nu}^2$, then $\pi(\theta|D)$ will be normal with mean $\frac{\theta/\nu^2 + \bar{\theta}/\bar{\nu}^2}{1/\nu^2 + 1/\bar{\nu}^2}$ and variance $\frac{1}{1/\nu^2 + 1/\bar{\nu}^2}$ where we again set $\theta$ to the point estimate based on data from prior studies and $\nu^2$ to the variance associated with that point estimate; using this mean and variance in place of $\theta$ and $\nu^2$ respectively in our PCES formulae (i.e., instead of simply the point estimate and the variance associated with it) yields the PCES under the assumption of a normal sampling distribution and an informative normal prior. However, regardless of the sampling and prior distributions assumed, the expected power of a replication study averaged over the prior is exactly calibrated to the desired level when the sample size for the replication study is set using the PCES corresponding to the assumed sampling and prior distributions.

Received April 3, 2014
Revision received February 10, 2015
Accepted February 24, 2015 ■